DOCUMENT RESUME

ED 263 212 TM 850 707

AUTHOR Perlman, Carole L.
TITLE Results of a Citywide Testing Program Audit in Chicago.
PUB DATE Apr 85
NOTE 10p.; Paper presented at the Annual Meeting of the American Educational Research Association (69th, Chicago, IL, March 31-April 4, 1985).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Descriptive (141)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Achievement Tests; Cheating; *Educational Testing; Elementary Education; Error Patterns; *Junior High Schools; Program Evaluation; *Public Schools; School Districts; Scores; Standardized Tests; Testing Problems; Testing Programs; Validity
IDENTIFIERS *Chicago Public Schools IL; Iowa Tests of Basic Skills; Test Reporting; Test Rescoring; Test Score Variance; *Test Security

ABSTRACT

Due to a concern regarding the validity of Chicago public elementary schools standardized test scores on the Iowa Tests of Basic Skills, a test audit was conducted. The audit was limited to 40 schools. Two schools were selected from each of the 20 administrative districts, with each school testing one seventh- and one eighth-grade class. These schools did not constitute a representative sample but were chosen because of possible anomalies. Other schools were retested to lend geographical balance to the sample and to provide an estimate of score change in a retest situation. There were a total of 23 suspect and 17 comparison schools. Nineteen of the 80 classrooms had median score declines that met or exceeded the set cutpoints. Of these classrooms, 17 were in suspect schools. It was suspected that exceeding the publishers' time limits was the most common way of increasing scores. Also, instances of altered answer sheets were detected using a computer program. However, district superintendents' investigations revealed no irregularities that would account for the score drops. Tentative plans for the spring 1985 testing provide for an increased emphasis on test and answer sheet security, adherence to time limits, and inservice training for testers. (PN)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED263212

RESULTS OF A CITYWIDE TESTING PROGRAM AUDIT IN CHICAGO

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

C. L. Perlman

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Carole L. Perlman

Chicago Public Schools

TM 850 707

Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, April 1985.

"I had a feeling something was wrong."

Henrietta K. Thomas on 1984 Chicago test scores, Chicago Tribune letters to the editor, July 31, 1984.

Background

In Chicago public schools, students in grades one through eight are tested each spring (usually by their classroom teachers) with a nationally standardized achievement battery. As is customary, the results are used for program evaluation and for monitoring systemwide progress. The results are also used for a variety of less defensible purposes.

Seventh grade scores are used to determine both which high school a student will attend and what courses he will be allowed to take. (Students who score below a fairly high cutpoint are required to take English and math classes that are either remedial or which only carry elective credit.) Eighth grade scores are used to determine whether a student will be allowed enter high school. Principals are evaluated on the degree to which their yearly objectives are met and there's always an objective that has to do with getting higher test scores. And although it's not official system policy, some principals base their teacher evaluations at least in part on their students' test scores. For all of these reasons, there is a lot of pressure to get high test scores.

Over the past years some concern has been expressed regarding the validity of the test scores. The concerned people tended to be research department staff as well as some community members and teachers. There was talk of teaching the test, of disregard for the publisher's time limits, and of altered answer sheets.

Although the research department staff had been troubled by this issue for some time (nobody wants to write evaluations based on questionable data), we were limited in what we could do about our concerns. First, the department has no line authority over the test givers. Second, there are nearly 500 elementary schools and rumors of testing irregularities tend to begin, "There's this school (I can't tell you the name of it, because my sister-in-law works there) where you wouldn't believe what they're doing with the test!" It was hard for us to spot anything other than schools with peculiar patterns of year-to-year gains (e.g., below-average sixth grade classes whose all-important seventh grade scores showed a median gain of more than two grade-equivalent years).

Given these facts, a test audit seemed like a logical thing to do. It became politically feasible for two reasons. First a citizens' group task force on schools recommended such a procedure. Second, board members and reporters were beginning to ask questions that could only be answered by an audit. It was clearly time to do something and the system administration consented to the audit.

Method

Because of limitations imposed by availability of staff, the audit was limited to 40 schools. Two schools were selected from each of 20 administrative districts with the system; in each school one seventh-grade class and one eighth-grade class was chosen for testing. Rooms with split grades or more than one or two special education students were not included in the retest.

The schools selected do not constitute a representative sample of the city's public schools. Instead some schools were chosen because of possible anomalies (unusual patterns of gains in previous years, unusually large

numbers of answer sheets ordered, or information received by the Department); these schools will be referred to as "suspect schools." Other schools (designated "comparison schools"), where there were no apparently aberrant results, were chosen to lend geographical balance to the sample and to provide an estimate of how much one would expect scores to change in a retest situation. There were a total of 23 suspect schools and 17 comparison schools.

The testers were either district guidance coordinators or teacher-certificated staff members of the Department of Research and Evaluation. The testers were experienced in giving tests and they received special inservice training for this project on the administration of the Iowa Tests of Basic Skills. Testers who routinely worked in particular districts were assigned to schools in other districts. Each tester was responsible for one or two schools and was instructed to follow the publisher's testing procedures precisely.

The retest took place the week immediately following the end of the citywide testing period; the schools involved were notified on the preceding Friday. The Iowa Test of Basic Skills Form 8 Reading Comprehension subtest was given to each of the two classes, usually on the same morning (Form 7 had been given during regular citywide testing). Students were told that their class had been specially selected along with other rooms throughout the city; that their scores would be returned to their school for use at the principal's discretion; that they should treat this test as though it were the actual citywide test; and that it was important for them to do their best. The test were administered to all students in the designated rooms who would normally participate in the citywide testing program (some special education students were exempt). Classroom teachers generally remained in the room during the testing period.

Results

Data analysis centered on answering two questions. First, how did the results from the retest compare with the citywide results? Second, how did the results from the suspect schools compare with those from the comparison schools?

Because of differences in student motivation, we expected that there would be a small decline between the citywide and retest scores. In fact, we did find that the median losses for the comparison schools were two months for seventh grade and one month for eighth graders*. The median losses for the group of suspect schools were five months for the seventh graders and four months for the eighth graders. For the suspect schools, the median change in grade-equivalent scores ranged from a gain of five months in one classroom to a loss of 11 months in four classrooms. For the comparison schools, the change in scores ranged from a gain of three months in one classroom to a loss of seven months in one classroom. Clearly, the suspect schools did much worse on the retest than the comparison schools.

Based on the standard error of the test and data from the comparison schools, decision rules were established to categorize schools as either having irregularities or not; cutpoints were set at seven months for seventh grade and six months for eighth grade. It was felt that smaller deviations might have a reasonable (although very small) likelihood of occurring by chance alone. (Conservative cutpoints were chosen to reduce the odds of misclassifying a school which may, by chance, have a large drop in scores). Nineteen of the eighty classrooms had median score declines that met or

*Our system routinely reports scores in terms of median grade-equivalents.

exceeded the cutpoints. Of these classrooms, 17 were in suspect schools (representing 37 percent of the classrooms in suspect schools), and two were in comparison schools (6 percent of the comparison classrooms).

Discussion

The results in some ways bring up more questions that they answer. First of all, there was a somewhat smaller than expected difference between the losses observed in the suspect and comparison groups. It's hard to tell exactly why. One possibility is of course, that our results do reflect the true nature of our populations. Perhaps we simply guessed wrong when we selected and categorized our schools.

For our comparison group, we chose schools that showed no unexpectedly high scores or unusual patterns of gains from 1982 to 1983. If there were fairly consistent irregularities over a period of several years, these irregularities would be hard to detect. Thus, we might have overlooked some cheating that may have been going on.

In the case of the suspect schools, we may have inadvertently chosen classrooms where cheating was least likely to occur. In our attempt to make life easier for the testers, we omitted from our study classrooms with more than one or two special education students; in fact, when we had a choice, we chose classrooms with no special education students. Since most of the mainstreamed special education students are learning disabled and since these students would probably assigned to "slower" classes, we probably managed to bias our results by choosing the best seventh-and eighth-grade classrooms in some of our schools. These are exactly the classrooms where it would be the least necessary to cheat, so it's possible that we may have underestimated the extent of cheating at some schools.

There's also the question of what caused the substantial declines that we did observe. Although we don't know for sure, we suspect that exceeding the publisher's time limits was the most common way of increasing the scores. It's easier and much more discreet than actually teaching the test, although that happens sometimes, too. Another method is providing the students with correct answers either during the test or by altering the answer sheets afterward. This is generally hard to detect, especially if only a few students' answer sheets are involved.

We were able to detect a number of instances of altered answer sheets by using a computer program that prints out item strings for each student in a class, percent of students getting each item correct, the expected percent correct based on the national norming sample, and percent of students choosing each distractor. We looked for strings of answers that nearly all students got right (especially if some of the questions were very hard) or items (particularly easy ones) that nearly everyone got wrong by choosing the same distractor. That last phenomenon is especially damning. You can argue that the kids got a hard item right because of superior teaching of the skill involved, but it's awfully hard to explain away an entire class of high scorers who chose the same wrong answer to an easy item.

Our school system does its own optical scanning so we were able to examine the citywide testing answer sheets for the rooms with funny-looking item statistics. The things to look for are unusually large numbers of erasures (random samples of the system's answer sheets at each grade were used as baselines) and a high proportion of answers that were changed from wrong to right. The average number of erasures on the citywide tests can also be compared with the number of erasures on the retest answer sheets. This

gives an indication of whether or not the citywide testing answer sheets had been tampered with.

One thing was clear: some changes had to be made in our system's testing procedures. The report concluded with recommendations for inservice training in correct testing procedures and new guidelines for maintaining test security.

Reactions

The initial audit report was released in July 1984. The schools with large declines between citywide testing and the retest were not named; the district superintendents were asked to conduct investigations of those schools; the results of the investigations were to be incorporated into a final audit report. Although the report was euphemistically worded (it never mentioned "cheating," but instead delicately alluded to "irregularities"), the newspaper articles about it still had headlines like "Reading score cheating" (Chicago Sun-Times) and "Lax teachers hit for 'high' test scores" (Chicago Tribune).

The final report was released the following winter. It listed the names of schools with substantial score declines on the retest (some classrooms showed median drops of as much as 1.1 years). It also gave the results of the district superintendents' investigations, namely that none of the investigations revealed any irregularties that would account for the drop in scores. This is a remarkable finding, especially since there was evidence that the answer sheets had been tampered with in at least four schools. The staff at those schools faulted the retest givers for not sufficiently motivating the students, but that argument didn't hold much water since the all of those schools had large declines over and above what we observed for our comparison schools, which served as a control as far as motivation was concerned.

One thing that most people agreed on was the need for an increased emphasis on test and answer sheet security, adherence to time limits, and inservice training for testers. Tentative plans for the spring 1985 testing provide for all of these and will not allow any fourth- through eighth-grade students to be tested by any teacher who gives them reading or math instruction. Principals will have to file testing schedules and security plans to their district superintendents.

Finally, and most important, the audit program will be continued and expanded. We feel that this is essential to ensure the validity of the scores from future testing programs.